

A Model of the Fresh Internet

Damien Lefortier, Liudmila Ostroumova, Egor Samosvat
Yandex, Moscow, Russia
{damien, ostroumova-la, sameg}@yandex-team.ru

Abstract

Previous models of the Web graph have highlighted some interesting properties, but have failed to describe the behavior of new content, especially how links to newly created pages appear. We experimentally study new Internet content using real-world data collected at Yandex (Russia's most popular search engine) and then we propose a new model of the Web graph, which reflects the behavior of such new content. We show through a set of experiments that this model realistically predicts the personalized PageRank and the diameter of new Internet content, something already existing models did not do. This model can be used for crawling, for example to define and tune crawl policies to improve the freshness of a search engine's index.

Keywords: Random graph models, Web graph models, fresh content, personalized PageRank.

1 Introduction

Numerous models have already been suggested to reflect and predict the growth of the Internet [4, 5, 12] but most of them are not suitable to describe the behavior of newly created pages. In particular, these models have problems describing how links to newly created pages appear and do not take into account the fact that new pages often tend to link to other new pages. E.g. in blog posts people usually cite recent blog posts of other people.

This paper, on the other hand, aims to solve the problem mentioned above and suggest a new model, which reflects basic properties of new Internet content.

One can divide Internet pages into two groups: *content pages*, which have information and *navigational pages*, which make it easier to find content pages. When a new content page appears on a site, it is usually referred to by some of the navigational pages. In this paper we call such navigational pages *hubs*. In order to model the appearance of new content on the Internet, we study the appearance of new links on hubs and also discuss the coverage obtained in this way.

Contributions of this paper are the following:

- We analyze the behavior of newly created pages using real-world data. We obtain that one needs to visit only 17% of all hubs to cover 90% of all fresh content. Also we analyze the evolution of links for fresh content pages and obtain some interesting properties.
- Based on these properties we suggest a new model, which behaves very similarly to the fresh Internet.
- We generate the model and compare its properties with the properties of the Web graph. First, we analyze the personalized PageRank and obtain that in both our model and in real data, fresh pages have a greater PageRank, as expected. Then we consider the evolution of diameter and show that in contrast to the small-world property [17, 18] the diameter of some parts of the Internet grows linearly with the number of pages.

The rest of the paper is organized as follows. In Section 2 we discuss prior related work. Section 3 describes the data that we used for our experiments. In Section 4, we experimentally study how new content appears on hubs before proposing a new model of the fresh Internet in Section 5 based on such experiments. Section 6 presents an empirical validation of our model in comparison with the Web graph. Finally, Section 7 concludes the paper and discusses possible applications.

2 Related work

One of the first attempts to propose a realistic mathematical model of Internet growth was made by Barabási and Albert in 1999 [3]. The main idea was to take into account the assumption that newly created pages often link to old popular pages. They defined a graph construction stochastic process, which is a Markov chain of graphs, governed by the *preferential attachment*. At each time step in the process, a new vertex is added to the graph and is joined to m different vertices already existing in the graph and chosen with probabilities proportional to their degree (measure of popularity). This model successfully explained some properties of the Internet like a small diameter and a power law degree distribution. Later many modifications to the

Barabási–Albert model have been proposed (Buckley–Osthus, Holme and Kim, Cooper–Frieze and others) aimed to more accurately depict these and other properties (see [1, 6] for details).

In context of our research, the main drawback of these models is that they pay too much attention to old pages and do not realistically explain how links pointing to newly created pages appear. The idea to combine preferential attachment with novelty factors was suggested in [13], where they considered a group of media sources, each of them reporting on a single topic (or thread) in one time period. Thread j can be chosen with probability proportional to the product $f(n_j)N(t - t_j)$, where n_j is the number of stories previously written about j , t is the current time, and t_j is the time when j was created. The function N is monotonically decreasing and therefore newer threads are more popular while the function f is monotonically increasing thus taking into account the idea of preferential attachment.

Our model is based on the same ideas. Each news page becomes out of date after some period of time and can not gain new links after that. The difference between our model and preferential attachment models is that each page has some inherent popularity and gains incoming links according to this popularity and not according to its degree.

3 Data

We consider two data sets that we used in our experiments to define and validate our model. The first one is a sequence of repeated crawls of a number of chosen hubs. The second is a snapshot of the data collected by one of Yandex’s continuous crawler called Orange [19], which is specifically designed for indexing frequently updated websites (see [14] for details on incremental crawling techniques).

Chosen hubs. For our research, we chose eight hubs from the Russian Internet: two of them are blogs (tema.livejournal.com, habrahabr.ru), four are news websites (lenta.ru, ria.ru, odnako.org, expert.ru), and two are forums (forums.drom.ru/general, rutracker.org), which are all quite popular in Russia. This dataset was created by crawling each chosen hub every 10 minutes for 1 week. During each crawl, we collected the number of newly discovered links on each hub and the number of outgoing links on each one of these new links. At the end, we obtained a dataset D_1 , which consists of 10,000 crawls and 30,000 discovered links overall.

Orange. This dataset was created by taking a subset of our crawler’s data centered on the Russian Internet. We obtained a dataset D_2 , which contains 3 billions documents and 6.5 billions links. For each document, we kept the history of the last 30 crawls (including the number of newly discovered links), the list of outgoing links from the last successful crawl and the document’s personalized PageRank [11]. Note that we used the 200,000 largest hubs as seed.

4 Analysis of new Internet content

In this section, we discuss and analyze the following properties of new Internet content: the appearance of new pages and the evolution of the links. We first analyze the behavior of outgoing links of hubs, then the distribution of links from and to content pages and finally discuss how to select hubs to cover most of the new content.

4.1 Evolution of outgoing links of hubs

Let us think of a hub as the creator of the pages it points to. In this way, we can identify the appearance of new links on a hub with the creation of new pages on the Internet. Furthermore, when the link to some page is deleted from the hub it usually means that the page somehow became outdated and that there will be less interest to it in terms of new incoming links.

Outgoing links count. The average number of outgoing links (or outdegree) of the hub is pretty much stable over time in general. Blogs and forums usually have constant outdegree because there is a fixed number of most recent posts on the main page. For news websites, this number may vary over time but does not deviate much from the average value. See Table 1 for the average number of outgoing links for our eight chosen hubs.

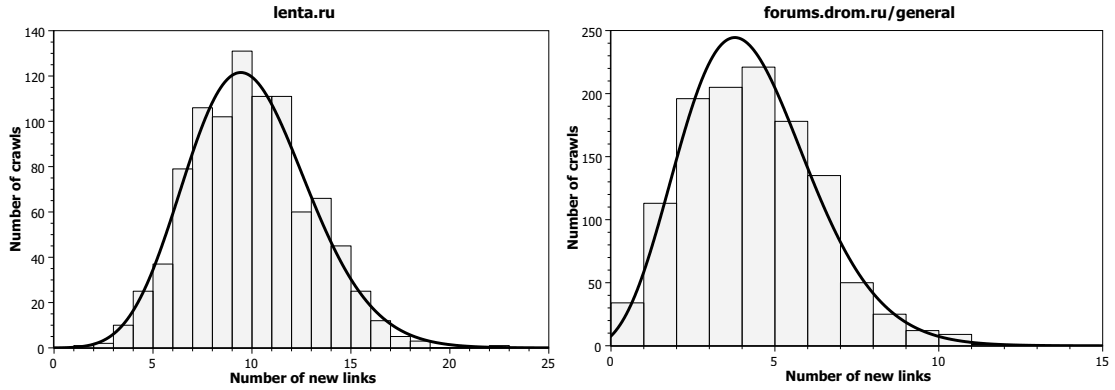


Figure 1: Number of new links discovered at each crawl versus theoretical prediction.

New links appearance. Let us now analyze how often new links appear on hubs. In [7, 8, 9], a Poisson process is used as an approximation model to describe Web page changes. We performed the same analysis for new links to show that new links appear on hubs according to a Poisson process¹, i.e. that new links are added to a hub randomly and independently with a fixed rate λ . We also show that each hub H_i has its own parameter λ_i .

We consider our eight chosen hubs and use the dataset D_1 (see Section 3) to verify that new links appear on these hubs according to a Poisson process. First, we estimated the parameter λ_i of each hub $H = H_i$ using a maximum-likelihood estimation, which gives us $\lambda = \frac{\sum x_j}{10}$ (we measure time in minutes), with x_j being the number of newly discovered links at the j^{th} crawl of H . The estimated intensities are in Table 1. Second, we used the Pearson's chi-squared test to verify that the number of newly discovered links at each crawl is the Poisson random variable with parameter $10\lambda_i$. So for each hub, we calculated the chi-squared test statistic X^2 , which is the normalized sum of squared deviations between observed and theoretical frequencies. We then compared X^2 to the critical value of no significance from the χ^2 distribution. The obtained p -values for our chosen hubs are between 0.15 and 0.5.

¹The time between each pair of consecutive events has an exponential distribution with parameter λ , which means that the time that the next event will occur has the probability density function $\lambda e^{-\lambda t}$ for $t > 0$. The number of events in the time interval $(t, t + \delta t]$ follows a Poisson distribution with parameter λ , which means that the probability that k new events occur during this time interval equals $\frac{e^{-\lambda \delta t} (\lambda \delta t)^k}{k!}$.

Figure 1 shows the observed frequencies versus theoretical ones. For better representation we used the following continuous approximation of the theoretical frequencies: $f(x) = e^{-10\lambda \frac{(10\lambda)^{x-0.5}}{\Gamma(x+0.5)}}$, where $\Gamma(x)$ is the gamma function.

	λ	l	m_{int}	m_{ext}	γ_{int}	γ_{ext}
lenta.ru	0.94	500	22	6	2.5	3
ria.ru	0.26	200	96	18	2.4	2.6
odnako.org	0.018	50	30	2	2.5	2.7
expert.ru	0.26	100	60	4	2.05	2.2
habrahabr.ru	0.12	65	35	4	2.05	2.5
tema.livejournal.ru	0.0033	20	2	6	2.4	2.1
forums.drom.ru/general	0.38	60	2	11	6.5	2.02
rutracker.org	0.014	60	4	3	2.4	2.4

Table 1: Properties of chosen hubs: λ — intensity of new links, l — average outdegree of a hub, m_{int} and m_{ext} — average internal and external outdegree of hub’s content pages, γ_{int} and γ_{ext} — parameter of internal and external indegree distribution.

4.2 Distribution of links of content pages

Outgoing links. In general, the text of a content page (e.g. a blog post, news, etc.) does not change much after it has been published and we can thus assume that outgoing links of content pages are static. Our hypothesis is that they point to other content pages according to the inherent quality of the target page. Let us now study the distribution of outdegrees of content pages of our chosen hubs.

We computed the outdegree distribution for each hub independently as shown in Figure 2. Note that we obtained quite chaotic distributions that definitely do not follow a power law. Therefore we computed the average page’s outdegree for each hub (see Table 1) as a good approximation of outdegree.

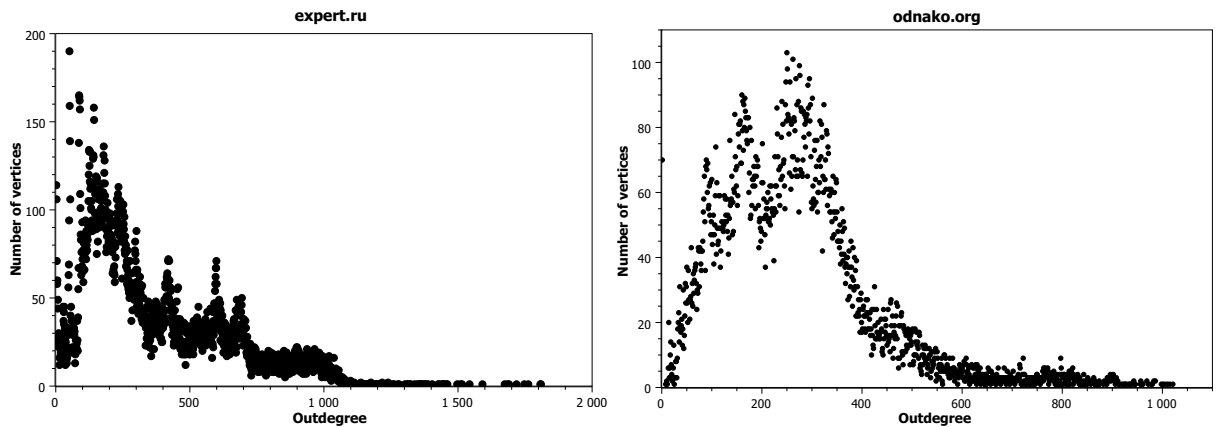


Figure 2: Outdegree distribution for two chosen hubs.

Incoming links. As mentioned in Section 1, one of the main problems of previous models is that they do not realistically explain how links to newly created pages appear. Our hypothesis is that a new page collects links before it becomes outdated and that the speed of this process depends on the inherent quality of the page; with better pages finally having a higher indegree. One can assume that the distribution of qualities of all pages created by the hub obeys some statistical law. Let us now check this assumption.

We calculated the internal ² indegree distribution for each hub independently and obtained that these distributions follow a power law with parameter γ_{int} that differs from hub to hub as shown on Figure 3. The obtained values are in Table 1.

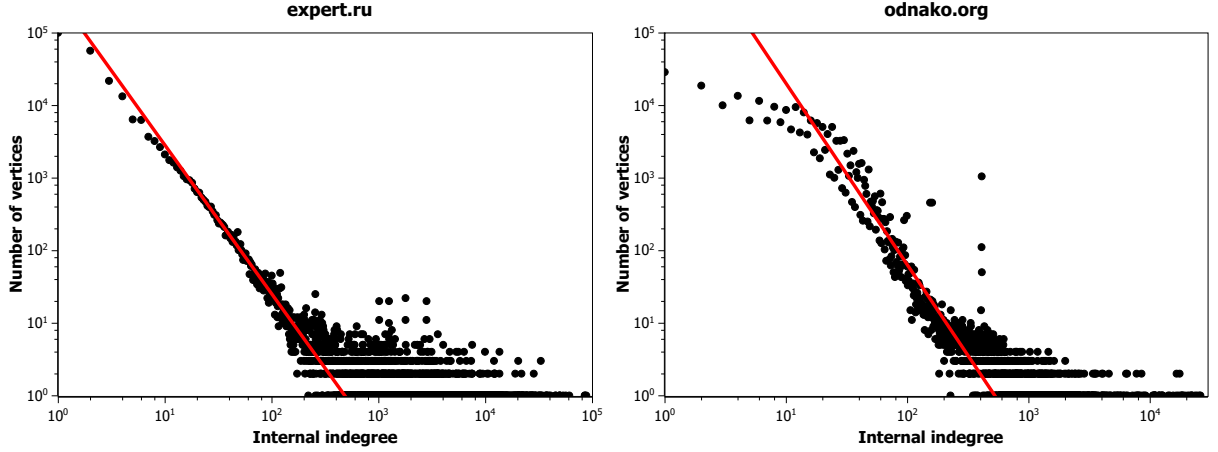


Figure 3: Internal indegree distribution for two chosen hubs.

We also calculated the external indegree distribution for each hub and obtained the same results (see Figure 4 and Table 1). Note also that γ_{int} and γ_{ext} can be different for one hub.

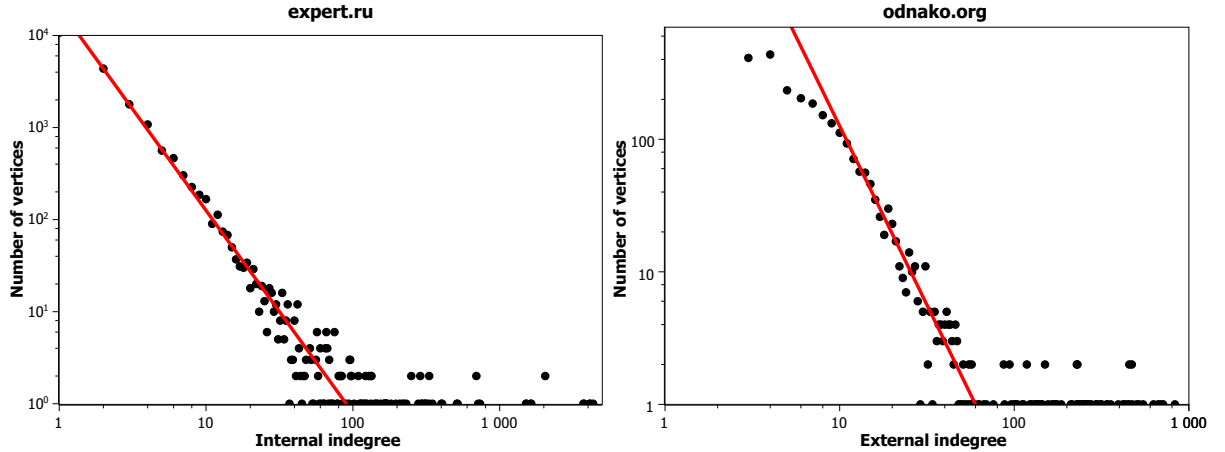


Figure 4: External indegree distribution for two chosen hubs.

²We say that a link is *internal* if it connects pages from the same site and *external* otherwise. One can thus split the overall indegree of a page into internal and external parts.

4.3 Cover size

Let us finally consider more hubs. Using the dataset D_2 (see Section 3), we created a set of hubs by taking the top 200 millions hubs according to their personalized PageRank. We chose this method for its simplicity, although there are better ways of doing this. For example, [16] suggested to use click information to find news, which then lead to pages pointing to more news, i.e. hubs. But our method is accurate enough for what we present here.

We computed the distribution of the parameter λ for all of these hubs using maximum-likelihood estimation: $\lambda_i = \frac{\sum x_{ij}}{\sum t_{ij}}$. Figure 5 shows the histogram of all parameters. Note that we obtain a heavy-tailed distribution.

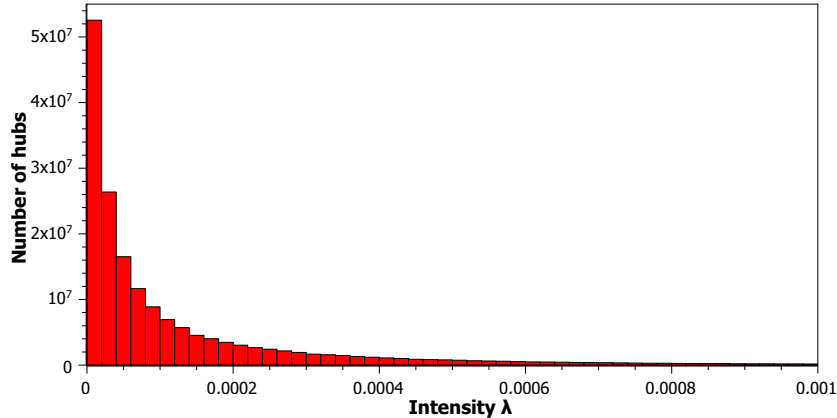


Figure 5: Parameters of the Poisson process for 200M hubs.

This parameter λ is the average number of new links, which appeared on a hub during one minute. Let us find the relative number of hubs needed to cover 90% and 95% of all new links. In order to do this, we use our representative set of 200M hubs. Let $S = \sum \lambda_i$ be the sum of all parameters and S_i be the sum of i larger parameters. We obtain that $S_{34M} > 0.9S$ and $S_{59M} > 0.95S$. Therefore we need only 17% of all hubs to cover 90% of all new links and 30% of hubs to cover 95%.

5 Description of the model

In this section, we propose a new model, which reflects the behavior of new Internet content. As we discussed in Subsection 4.3, most new content is covered by a few hub pages. Suppose that we have n such hubs H_1, \dots, H_n . We call the pages created by a hub the ‘children’ of this hub. Each hub H_i has the following four characteristics: the intensity of new links appearance λ_i , the number of outgoing links l_i , the children’s outdegree m_i , and the parameter of power law of the children’s quality γ_i .

At the beginning of the process we have the hubs H_1, \dots, H_n with no links. Then new links appear according to a Poisson process. Poisson processes for different hubs are independent.

Suppose that at some point we have a graph G , consisting of the hubs H_1, \dots, H_n and other pages. Each hub $H = H_i$ has l outgoing links (or outlinks) to some pages p_1, \dots, p_l (children of this hub). When a new outlink on a hub H appears then a new page p is created and the oldest

outlink of H is deleted. The quality of p is set to some random number $P = \frac{1}{r^{1/(\gamma-1)}}$, where r is the uniform random variable on $[0, 1]$. Also m links from p to other pages currently present on any of the hubs appear. The probability to choose each page is proportional to its quality.

This process is natural because when a new page appears on the Internet, it has some inherent popularity and although it has no links at first, it can become very popular and gain incoming links rapidly. This is the main difference between the idea of preferential attachment and our model: in preferential attachment models, older vertices are usually more popular because newly created pages have no incoming links and are thus unpopular in the sense of preferential attachment.

In order to verify our model, we generate several graphs using the algorithm defined above. For simplicity, we take $n = 1$, i.e. generate graphs with one hub. To make them realistic we use parameters from the Table 1. In particular we simulate *odnako.org* and *ria.ru*.

As shown in Table 1, for each hub the average internal and external outdegrees of content pages differ. Internal and external quality (γ_{int} and γ_{ext}) are also different. In our experiments, we use the parameters m_{int} and γ_{int} for the outdegree and the quality of children because we model only one hub. It is natural to use m_{int} and m_{ext} for the internal and external children's outdegrees when we model a graph with several hubs. Also when a new page appears on a hub, it can have an internal and external quality γ_{int} and γ_{ext} , so that this new page will gain new internal links according to its internal popularity and new external links according to its external popularity.

Basic feasibility. In order to ensure that our model can actually predict the behavior of new Internet content, we need to verify that the internal indegree distribution of graphs generated using this model follows a power law as for the Web graph (see Subsection 4.2). As an experiment, we computed this distribution for random graphs that we generated according to our model as described above and as we can see in Figure 6, this distribution does follow a power law with expected parameters. In the next section, we investigate other properties of our model in comparison with the Web graph to get further validation.

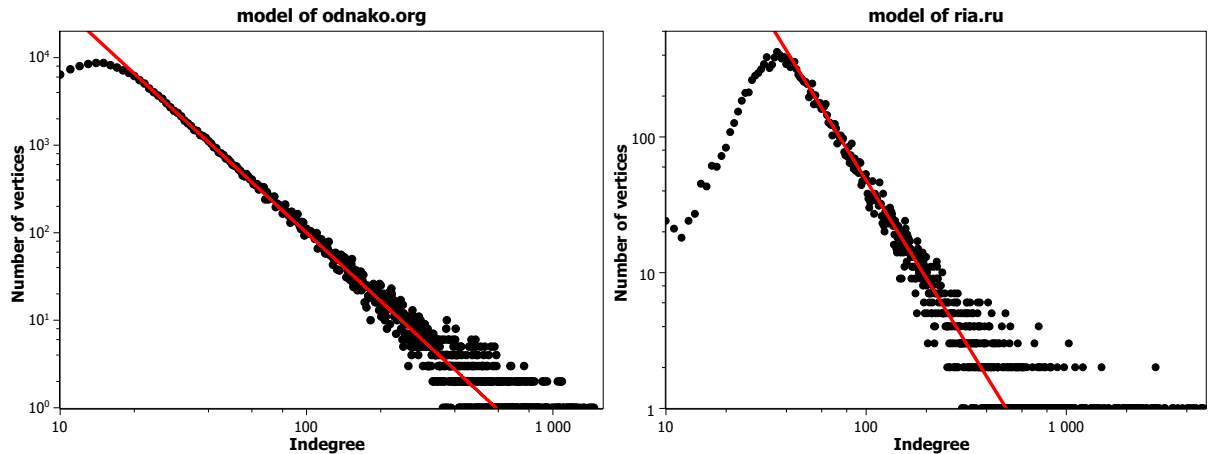


Figure 6: Indegree distribution for generated graphs.

6 Empirical validation

In this section, we consider a set of important properties of our model and investigate these properties for the Web graph as an empirical validation. We generated random graphs according to our model as described in Section 5 and then compared the following properties between these graphs and our real-world dataset D_2 from Section 3: personalized PageRank (see Subsection 6.1) and diameter (see Subsection 6.2).

The behavior of these properties is not realistically predicted by the numerous random graphs models that have been proposed to reflect other important quantitative and topological aspects of real-world networks like the Internet, including preferential attachment models (see Section 2). We show that our model realistically predicts both of these properties.

6.1 Personalized PageRank

One important aspect of the Internet is that newly created web pages can be extremely popular, while older pages can be out of date even though they can have many incoming links. We use personalized PageRank (with hubs as seed) to show that in our model, newly created pages tend to have a higher rank than older ones. Then we do the same analysis for real web graphs.

Case of our model. We computed the personalized PageRank of random graphs generated according to our model and as shown respectively in Figure 7 and 8, this rank decreases in time (when indegree is fixed) and grows linearly in degree (for pages created in a small time interval), which indeed means that in our model new content tends to have a higher rank than older content. Note that we used the hub as seed when computing the personalized PageRank of each graph.

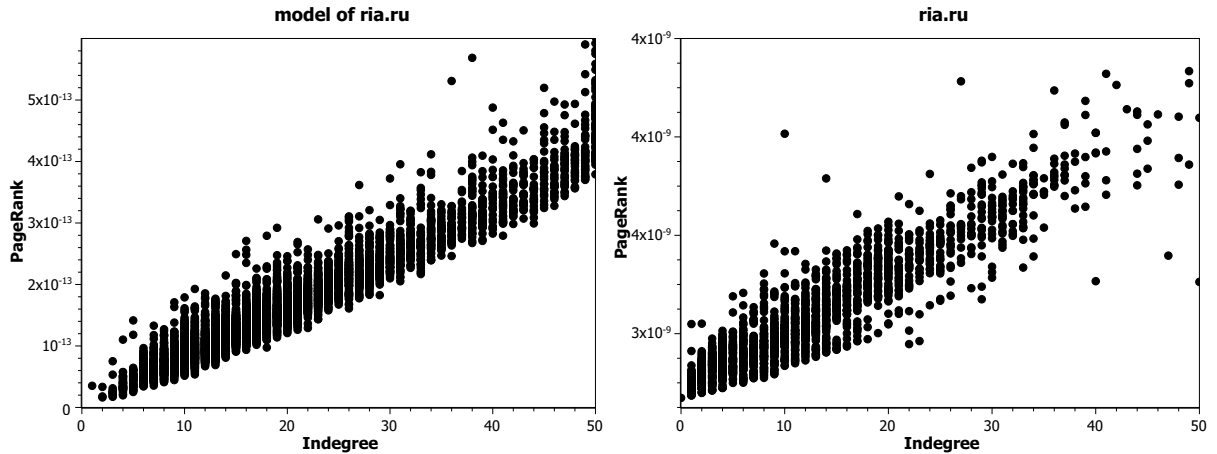


Figure 7: Personalized PageRank depending on degree for ria.ru versus generated graph.

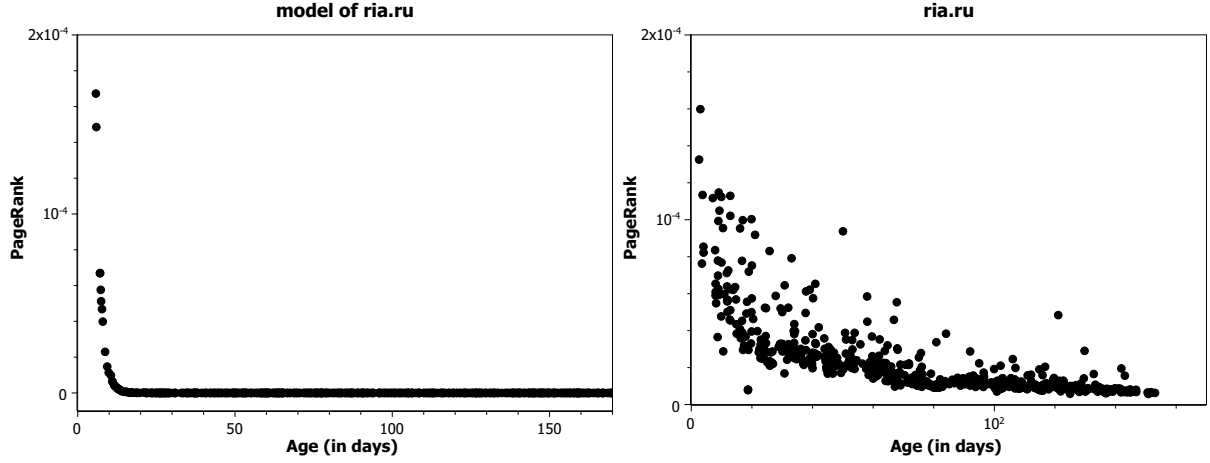


Figure 8: Personalized PageRank depending on time for ria.ru versus generated graph.

Case of web graphs. Using D_2 , we computed the web graph of each chosen hub's host by considering all pages within this host and all links, whose source and destination belong to this host and which are neither navigational nor advertisement. For example, the graph for ria.ru contains approximately 350,000 pages and 110 million links. Then we computed the personalized PageRank of these graphs and as shown in Figure 7 and 8, this rank is also decreases in time and grows linearly in degree, which means that our model is an accurate representation of new Internet content in that aspect.

6.2 Diameter

Interestingly, most models that have already been proposed exhibit the small-world property [15, 17, 18], which means that most nodes can be reached from every other by a small number of hops. We use the diameter to show that in our model, this property is not verified. Then we do the same analysis for real web graphs.

Case of our model. We computed the diameter of random graphs generated according to our model and as shown in Figure 9, the diameter of such graphs is linear in the number of vertices, which means that the small-world property does not stand for our model.

Case of web graphs. Using D_2 , we computed the web graph of each chosen hub's host as described in Subsection 6.1. Then we computed the diameter of these graphs and as shown in Figure 9, this diameter is linear in the number of vertices. This interesting property of the Web graph was not presented before and the main difference with previous studies is that we do not consider navigational links. It also reflects the fact that when a content page becomes outdated, then it gets harder to find from the hub. Our model does predict this property realistically, but the slope of the line is steeper for the model. One of the reasons may be the following: we need to filter the data more careful to get exactly what we want to study.

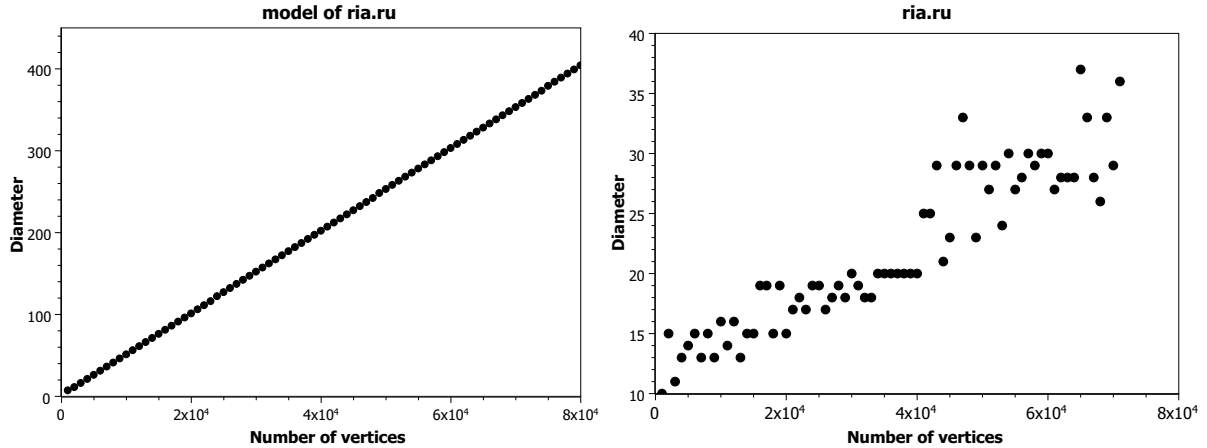


Figure 9: Diameter depending on the number of vertices for ria.ru versus generated graph.

7 Conclusion

We analyzed the behavior of newly created pages on hubs by using data collected by crawling the Web at Yandex and experimentally obtained properties of new Internet content. Then we used these experimental results to introduce a new model of the Web graph, which reflects the behavior of the new content. Finally, we empirically validated this model by comparing it to the Internet through further experiments. Our conclusion is that our model realistically predicts the personalized PageRank and the diameter of new Internet content, which was not the case with previously proposed models thus making a step forward towards a better understanding of Internet growth.

Applications. Web graph models can be used to define and tune crawl policies. For example, [9] proposed a Poisson process as the change model for data sources where each element changes at its own rate λ_i and suggested a refresh policy to maximize the freshness of the local copy with the two following metrics to measure the quality of such algorithm: freshness and age. In the same way, one can in principle maximize the freshness of a search engine's index by crawling pages, whose changes have been predicted by our model.

This idea of crawling pages which generate many new pages was suggested in [10]: old pages are ordered according to the number of new links they produced in the past and then top pages are revisited by the crawler. Using our model, one can get more precise predictions about when to revisit each page from its crawl history.

This work can also be used to estimate the freshness of such index by running crawling algorithms on random graphs generated according to our model and then looking at the delays between the discovery of each page and the actual time it appeared. Such delays are usually really hard to estimate because the exact time most pages appear is unknown and that can be incredibly useful to understand the quality of crawling algorithms regarding freshness.

References

- [1] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, *Reviews of modern physics*, vol. 74, pp. 47–97, 2002.
- [2] K. Avrachenkov, D. Lebedev, PageRank of Scale-Free Growing Networks, *Internet Mathematics*, vol. 3(2), 207–231, 2006.
- [3] A.-L. Barabási, R. Albert, Emergence of Scaling in Random Network, *Science* 286, 509, 1999.
- [4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics, *Physics reports*, vol. 424(45), pp. 175-308 (2006)
- [5] A. Bonato. A Survey of Models of the Web Graph. In: A. López-Ortiz and A. Hamel (Eds.): CAAN 2004, LNCS 3405, pp. 159–172, 2005.
- [6] B. Bollobás, Mathematical results on scale-free random graphs, *Handbook of Graphs and Networks*, pp. 1-34, 2003.
- [7] B.E. Brewington, G. Cybenko, How dynamic is the web? *Proceedings of the Ninth International World-Wide Web Conference*, 2000.
- [8] B.E. Brewington, G. Cybenko, Keeping up with the changing web, *IEEE Computer* 33, 5, 52–58, 2000.
- [9] J. Cho, H. Garcia-Molina, Effective page refresh policies for Web crawlers, *ACM Transactions on Database Systems (TODS)*, Vol. 28, No. 4., pp. 390-426, 2003.
- [10] A. Dasgupta, A. Ghosh, R. Kumar, C. Olston, S. Pandey, A. Tomkins, The discoverability of the web, *International conference on World Wide Web*, 421–430, 2007.
- [11] G. Jeh, J. Widom, Scaling Personalized Web Search, *Technical Report*, Stanford, 2002.
- [12] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, E. Upfal, *Web as a graph*, *Proc. PODS 2000*, pp. 1-10, 2000.
- [13] J. Leskovec, L. Backstrom, J. Kleinberg, Meme-tracking and the dynamics of the news cycle, *Proc. ACM SIGKDD*, 497–506, 2009.
- [14] C. Olston, M. Najork, Web Crawling, *Foundations and Trends in Information Retrieval*, Vol. 4, No. 3., pp. 175-246, 2010.
- [15] O. Sandberg, I. Clarke, The Evolution of Navigable Small-World Networks, 2008.
- [16] Y. Wang, Y. Liu, M. Zhang, S. Ma, News Page Discovery Policy for Instant Crawlers, *AIRS*, pp. 520-525, 2008.
- [17] D.J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton University Press, Princeton, NJ, 1999.
- [18] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, *Nature* 393, 440–442, 1998.
- [19] <http://company.yandex.com/technologies/searchindex.xml>